Probabilistic Guarantees for Abductive Inference

Anonymous authors
Paper under double-blind review

Abstract

Abductive reasoning is ubiquitous in artificial intelligence and everyday thinking. However, formal theories that provide probabilistic guarantees for abductive inference are lacking. We present a quantitative formalization of abductive logic that combines Bayesian probability with the interpretation of abduction as a search process within the Algorithmic Search Framework (ASF). By incorporating uncertainty in background knowledge, we establish two novel sets of probabilistic bounds on the success of abduction when (1) selecting the *single* most likely cause while assuming noiseless observations, and (2) selecting *any* cause above some probability threshold while accounting for noisy observations. To our knowledge, no existing abductive or general inference bounds account for noisy observations. Furthermore, while most existing abductive frameworks assume exact underlying prior and likelihood distributions, we assume only percentile-based confidence intervals for such values. These milder assumptions result in greater flexibility and applicability of our framework. We also explore additional information-theoretic results from the ASF and provide mathematical justifications for everyday abductive intuitions.

1 Introduction

Imagine a patient visits a doctor because of a persistent cough, fever, and shortness of breath. As the doctor considers these symptoms and the prevalence of certain illnesses in the area, the doctor may hypothesize that the patient has pneumonia. This is an example of abductive reasoning, or *abduction*.

Abduction is the process of finding the best causal explanation given some observed effects. Abductive reasoning can be categorized into strategies that can generate new hypotheses, known as *creative abduction*, and those that select the best candidate given a set of possible explanations, known as *selective abduction* (Schurz, 2007). We focus on selective abduction, which can be formalized with Bayesian Decision Theory (Romeijn, 2013). Given observation(s) O, we select a hypothesis C_i from a finite set of hypotheses C. Per Bayesian probability, we denote $\Pr(C_i|O)$ as the *posterior*, where the most probable cause is that with the highest posterior. By Bayes' theorem,

$$\Pr(C_i|O) = \frac{\Pr(O|C_i)\Pr(C_i)}{\Pr(O)}.$$

However, during the hypothesis selection process, the relevant observations $\Pr(O)$ remain constant. Thus, the relevant form of Bayes' theorem becomes

$$\Pr(C_i|O) \propto \Pr(O|C_i)\Pr(C_i).$$

To perform selective abduction, one simply chooses the hypothesis whose likelihood and prior have the greatest product. At first glance, it may seem that selecting a general cause (i.e., one that is likely and can produce many effects) will maximize the posterior due to its high prior. However, a general cause will have a smaller likelihood for any specific outcome, which decreases its posterior.

Abduction accompanies induction and deduction as one of three forms of logical reasoning (Rodrigues, 2011; Peirce et al., 2017). In supervised machine learning, inductive and abductive processes serve as the underlying logic behind model training and application (see Figure 1) (Mooney, 2000). While both inductive and abductive reasoning are applied ubiquitously in the field, inductive reasoning is currently the more well-understood process; we have already gained a theoretical understanding of inductive accuracy (Dietterich, 1989; Kietz, 1993; Cummings et al., 2016; Garg et al., 2021; Cosentino et al., 2022). However, to our knowledge, there currently exist no formal quantitative frameworks with accuracy bounds for abductive reasoning.

In a broader context, artificial intelligence experts such as Erik Larson argue that obtaining a quantitative theory of abduction is a necessary step towards bridging machine and human intelligence. Abduction, more specifically creative abduction, encapsulates human intuition or "guessing" capability lacking in current models. Larson describes machine understanding of abductive reasoning as the central "blind spot" of artificial intelligence:

> *"Abductive inference is required for general intelligence, purely inductively inspired techniques like machine learning remain inadequate…The field requires a fundamental theory of abduction." (Larson, 2021)*

Our work primarily aims to (1) provide currently lacking accuracy bounds for abductive reasoning and (2) serve as a preliminary version of this "fundamental theory of abduction" needed for abductive machine understanding. We propose a general probabilistic framework for *selective* abduction built from Bayesian Decision Theory (Berger, 2013) (detailed in Section 3), serving as a jumping off point for future work on creative abduction. Through this Bayesian framework, we first derive upper and lower probabilistic bounds of abductive accuracy when assuming underlying q-percentile uncertainty bounds of prior and likelihood probabilities for each cause (Section 4). This first set of accuracy bounds treats successful abduction as choosing the *single* true hypothesis assuming the selection of the single highest posterior. We then extend this by reframing abduction as a search process within the Algorithmic Search Framework (ASF) (Montañez, 2017), which lets us describe and bound the probability of selecting *any* hypothesis with a posterior probability above a certain threshold while accounting for noisy observations (Section 5.1). Lastly, in addition to deriving bounds on abductive accuracy, we apply the framework to quantitatively justify common-sense heuristic abduction (Section 5.2, 5.3).

2 Related work

We review applications of abductive logic in machine learning and artificial intelligence, and survey existing abductive frameworks and current literature on Bayesian inference.

2.1 Logic in Machine Learning

Peirce popularized abduction along with two other components of logical inference: induction and deduction (Shanahan, 1986). Induction, inferring causal relationships from data, is the primary task of machine learning (Mooney, 2000). Inductive logic is core to the *training* process, where labeled examples are used to derive generalized relationships within a trained model. Deductive and abductive logic are then employed within machine learning's underlying inductive framework by applying the relationships derived through inductive training (Bergadano et al., 2000). While deduction is the logical process behind data generation (namely, selecting a class (cause) from a trained model with known/assumed causal relationships to generate feature data (observations)), abduction is the logic underlying the task of assigning a class label (cause) to unlabeled data (observations) using a trained model that embeds known/assumed causal relationships (see Figure 1).

Inductive and abductive processes are particularly common in machine learning. One can map induction to the process of training, where input-output relationships are learned, and abduction to classification, where one uses known relationships to infer likely generating causes. Table 1 gives a schematic outline of logical inference (Bergadano et al., 2000) and machine learning.

Through this perspective, we see that machine learning applies abductive logic through the inference of trained models. For example, machine learning emulates the abductive reasoning of spam detection and medical diagnosis by applying trained algorithms to unlabeled data (i.e., text from emails or radiology scans). However, model inference is just one of many applications of abduction in machine learning. Our work

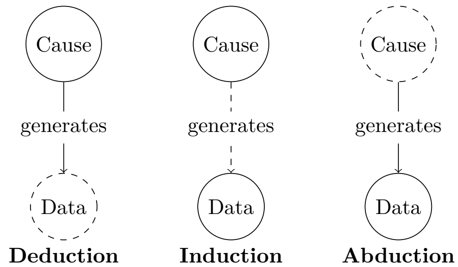

Figure 1: Three methods of inference. The dotted lines show which part of each process is being inferred.

addresses the theoretical limits of the success of abductive reasoning generalizable to applications such as these.

Table 1: Schematic outline of the processes of inference in supervised machine learning (Bergadano et al., 2000).

Logical Inference	**Machine Learning**
Induction: $P(a)$ $\therefore \forall x P(x)$	Training: $(x^1, y_1), ..., (x^n, y_n)$ $\therefore f : \mathcal{X} \rightarrow \mathcal{Y}$
Abduction: $Q(a)$ $P(x) \rightarrow Q(x)$ $\therefore P(a)$	Classification: $\mathbf{x}^m$ $y_m = f(\mathbf{x}^m)$ $\therefore y_m$

2.2 Applying Abduction in Machine Learning

In addition to its synonymy with the higher-level logic of model inference, abductive logic is central to several common machine-learning processes.

Abduction is the underlying logic of Bayesian networks, a system used for a broad range of tasks in machine learning, including clustering, supervised classification, multi-dimensional supervised classification, anomaly detection, and temporal modeling (Mihaljević et al., 2021). Bayesian networks are especially useful for creating algorithms that perform decision-making in domains with inherent uncertainty (Mihaljević et al., 2021). Thus, such networks are widely used in criminology and prognosis, diagnosis, and prescription in healthcare (Song et al., 2021).

Additionally, *maximum likelihood estimation* (MLE) and *maximum a posteriori* (MAP), common model training techniques, apply abductive reasoning to optimize model parameters.[1] Analogizing training data D as observations and a possible model parameterization $\mathbf{w}$ to a possible cause, these methods optimize model parameters by maximizing the posterior probability $\Pr(\mathbf{w}|D)$ through Bayes' theorem with likelihood $\Pr(D|\mathbf{w})$ and prior $\Pr(\mathbf{w})$ probability distributions (Bishop, 2006).

Abductive reasoning is also prevalent in areas such as relational learning and computer vision. Relational learning, where data is represented through its relationships or closeness with other data (Buhmann et al., 2011), uses abduction to guide search and generate missing input data (Bergadano et al., 2000). In computer vision, abductive reasoning can be combined with convolutional neural networks (CNNs) to solve spatial-temporal reasoning tasks (Zhang et al., 2021). By integrating an abductive reasoning system with CNNs, the resulting system can also potentially be used for image segmentation (Rafanelli et al., 2023). Additionally, the integrated system contributes to "explainable AI," an area that incorporates understandable reasoning into black-box models (Rafanelli et al., 2023).

[1]Note that training remains an inductive process on a larger scale; MLE and MAP apply abductive logic within training steps.

2.3 Formalizations of Abduction

Various formalizations of abduction have been explored in symbolic artificial intelligence literature (Paul, 2000). These include set-cover-based approaches (Allemang et al., 1987), which involve the requirement of complete causal relationships and entail selecting a subset of hypotheses from a larger set (referred to as hypothesis assembly), and knowledge-level approaches, which propose explanations based on beliefs (Levesque, 1989). Another notable approach is Abductive Logic Programming (ALP) (Kakas et al., 1992; Alberti et al., 2008; Raghavan & Mooney, 2010), where inferences are represented as entailments from a prior bank of knowledge to the veracity of specific causes leading to observed phenomena.

Probabilistic Horn Abduction (Poole, 1991; Ng & Mooney, 1991) is a logic programming language extending Prolog. This framework combines exact probabilities of hypotheses with Bayes' theorem to generate posterior probabilities built from multiple observations. Unlike our proposed framework, Probabilistic Horn Abduction assumes exact prior and likelihood probabilities and does not incorporate underlying confidence ranges for likelihood and prior distributions (Poole, 1991).

A recently developed framework applying stochastic mathematical systems (SMSs) can model abduction by representing the reasoning process as stochastic mathematical systems, where the human reasoner SMS generates hypotheses and an oracle SMS evaluates their validity based on explanatory power and evidence Wolpert & Kinney (2024). But like Probabilistic Horn Abduction, it also does not account for the uncertainties in underlying distributions.

All these methods lack probabilistic guarantees for the correctness of the abductive inferences, and uncertainties associated with the process remain unquantified. To mitigate this gap, our approach integrates formal machine learning frameworks, which allows for more precise quantification of the uncertainties involved in abductive inferences.

2.4 Bayesian Inference

Bayesian inference is the foundation of our framework, where we derive accuracy bounds based on assumed q_p and q_l-confident intervals for underlying prior and likelihood distributions, respectively. Such intervals act as parameters of how confident one is in their background knowledge of either causal relationships (indicated by q_l and the size of the likelihood confidence interval) or knowledge of what is generally common in the world (indicated by q_p and the size of prior confidence interval). This allows for more flexibility in what the framework can represent.

Bayesian inference estimations and bounds are well-explored in the literature, with numerous known methods of deriving accuracy bounds for inference of specific algorithms or tasks (Yekutieli, 2012; Pati et al., 2018; Chérief-Abdellatif et al., 2019; Alroobaea et al., 2020; Audibert, 2009; Zhang et al., 2021; Alquier & Ridgway, 2020; Ferguson et al., 1992; Cox, 1993; Alvarez et al., 2014). There are fewer existing methods of deriving general bounds for Bayesian inference through set techniques such as multi-valued mapping (Dempster, 1968) or prior measure intervals (Dempster, 1967). However, to our knowledge, there does not yet exist a method of deriving Bayesian inference bounds based upon specific prior and likelihood confidence intervals with certain probabilities q_l that affords our framework its versatility.

Our work is the first to leverage the ASF (Montañez, 2017) to construct a formalization of abduction or abduction by Bayesian inference. Unlike established abductive logic frameworks (Poole, 1991; Ng & Mooney, 1991; Poole, 1993), the ASF accounts for noisy observations – observation that may not fully reflect "true" events, perhaps representing a faulty observer or data pipeline with which observations are processed. The framework makes very few assumptions of given information resources, F, which (in the case of abduction) embeds observation data. Such data is abstracted as binary strings, with no conditions placed on what form the binary strings take, only that we have functions available to extract feedback from the strings for individual search queries. Thus, with no restrictions placed on the information resources, the ASF accommodates both noisy and noiseless observations. To our knowledge, there are no abductive or general inference bounds with this specific property. Existing work has only analyzed the correlation of real dataset noise with the accuracy of Bayesian inference for specific algorithms, assuming specific data qualities (An et al., 2012).

3 Preliminaries

We formalize the fundamental building blocks of abduction, causes and observations, as vectors. The vectorization of such outcomes allows the formalization of posterior, likelihood, and prior probabilities as distributions over a vector space. We then formalize the likelihood and posterior uncertainty intervals on which the abductive search process relies.

3.1 Vectorizing Observations

We formalize observations as vectors, where each scalar component corresponds to a specific *observation feature* that indicates the presence of a certain observed outcome. For simplicity, we consider only binary observation features indicating the existence or non-existence of an outcome. For example, suppose you swallow a pill and observe your headache disappearing. An observation vector might consist of two binary features representing (1) "Did you swallow a pill?" and (2) "Did the headache go away?" where "1" answers "yes" and "0" answers "no" to the respective question. If we observe that we swallowed a pill and the headache disappeared, the respective observation vector would be $\langle 1, 1\rangle$, whereas if the headache disappeared without taking a pill, the respective observation vector would be $\langle 0, 1\rangle$.

We denote the space containing all possible observation vectors as the *observation space* $\mathcal{O}$, defined as follows.

Definition 3.1. ($\mathcal{O}$) Let $\mathcal{O}$ denote the vector space with discrete topology containing all binary-featured observation vectors whose components indicate the existence or non-existence of some observed outcome. $\mathcal{O}$ contains $2^{\dim(\mathcal{O})}$ possible observation vectors.

Since any outcome must strictly occur or not occur, it follows that the set of possibilities within $\mathcal{O}$ is mutually exclusive and collectively exhaustive.

Proposition 3.1. *The set of possible outcomes represented as vectors within $\mathcal{O}$ is mutually exclusive and collectively exhaustive.*

3.2 Vectorizing Causes and Likelihood Probability Mass Functions

A cause C_i has some probability of instigating any possible observation vector $\mathbf{x} \in \mathcal{O}$, inducing a conditional probability mass distribution (i.e., likelihood function) $\Pr(\mathbf{x}|C_i)$ over all observations $\mathbf{x} \in \mathcal{O}$. Note that every observation $\mathbf{x} \in \mathcal{O}$ is disjoint (Proposition 3.1.2), and we assume exactly one observation vector is produced and observed.

Following the earlier example, the likelihood distribution over the observation space for the cause "you took aspirin" expresses the probability that, *assuming* aspirin was taken, phenomena $\mathbf{x} \in \mathcal{O}$ would follow. Knowing that (1) aspirin is typically taken by pill and (2) that aspirin typically cures most headaches, the likelihood probability distribution of cause "aspirin" over the observation space with dimensions {"Pill taken?", "Headache relieved?"} might place the following relative probability mass on the four possible observation vectors in $\mathcal{O}$:

Table 2: Example likelihood distribution for effects of aspirin.

Pill taken?	Headache relieved?	$\mathbf{x}$	$\Pr(\mathbf{x}\|\text{aspirin})$
no	no	$\langle 0, 0\rangle$	0.05
no	yes	$\langle 0, 1\rangle$	0.10
yes	no	$\langle 1, 0\rangle$	0.15
yes	yes	$\langle 1, 1\rangle$	0.70

For each of the four possible outcomes, we can formalize how likely each is relative to the others. Assuming that exactly one of the observation vectors must occur, we know that the probabilities for each collectively must sum to one. Considering all the possible ways there are to assign probabilities to a collectively exhaustive and mutually exclusive set of options forms a mathematical simplex, $\mathcal{S}$. For k observation features or $\dim(\mathcal{O}) = k$, simplex $\mathcal{S}$ forms a continuous $2^k - 1$ dimensional hyperplane containing all possible "cause vectors", each corresponding with some likelihood probability mass function over the 2^k observation vectors in $\mathcal{O}$. Each scalar component of a "cause" vector $\mathbf{c} \in \mathcal{S}$ denotes how much probability mass is placed on a corresponding observation vector in $\mathcal{O}$.

Ensuring that every cause $\mathbf{c} \in \mathcal{S}$ corresponds to a valid probability mass function on $\mathcal{O}$ requires the following two properties: (1) the simplex is bounded within $[0, 1]$ on every dimension such that no $\mathbf{c} \in \mathcal{S}$ holds a component that indicates an invalid probability, and (2) the sum of all components of a cause vector is upper bounded by 1.

3.3 Defining Posterior Confidence Bounds

During the decision-making process, the posterior $\Pr(\mathbf{c}|\mathbf{x})$ is calculated by scaling the likelihood $\Pr(\mathbf{x}|\mathbf{c})$ by the prior $\Pr(\mathbf{c})$ and inversely by the evidence $\Pr(\mathbf{x})$, per Bayes' theorem. When comparing different posterior probabilities for the same observation $\mathbf{x}$, the product $\Pr(\mathbf{x}|\mathbf{c})\Pr(\mathbf{c})$ becomes the only relevant factor as $\Pr(\mathbf{x})$ scales all alternatives equally, namely,

$$\Pr(\mathbf{c}|\mathbf{x}) = \frac{\Pr(\mathbf{x}|\mathbf{c})\Pr(\mathbf{c})}{\Pr(\mathbf{x})} \propto \Pr(\mathbf{x}|\mathbf{c})\Pr(\mathbf{c}).$$

Note that, for simplicity, the word "posterior" will henceforth refer to the product $\Pr(\mathbf{x}|\mathbf{c})\Pr(\mathbf{c})$.

In real life, we often lack these exact likelihood and prior distributions. Instead, we may estimate such probabilities through numerical techniques, including asymptotic estimations, Monte Carlo methods, numerical integration, and various sampling methods (Tierney, 1994; Chib, 1996; Levine & Casella, 2001). Other distribution estimation methods include smoothing and reduction methods, and Markov chain algorithms can be further used to combine estimation methods (Tierney, 1994). Thus, to account for uncertainty, we estimate likelihood, prior, and posterior probabilities through confidence intervals.

Rather than using one function mapping each vector in the cause simplex to a single likelihood distribution, we define two functions denoting the upper bound likelihood, $l_U(\mathbf{c}, \mathbf{x}) : \mathcal{S}, \mathcal{O} \to [0, 1]$, and lower bound likelihood, $l_L(\mathbf{c}, \mathbf{x}) : \mathcal{S}, \mathcal{O} \to [0, 1]$ of the q_l-percentile likelihood uncertainty interval, where $l_U(\mathbf{c}, \mathbf{x}) \geq l_L(\mathbf{c}, \mathbf{x})$. The prior q_r-percentile uncertainty interval is similarly represented through an upper and lower bound $r_U(\mathbf{c}) : \mathcal{S} \to [0, 1]$ and $r_L(\mathbf{c}) : \mathcal{S} \to [0, 1]$ (respectively) where $r_U(\mathbf{c}) \geq r_L(\mathbf{c})$.

The upper and lower bounds of the posterior confidence interval bounds, $p_U(\mathbf{c}, \mathbf{x}) : \mathcal{S}, \mathcal{O} \to [0, 1]$ and $p_L(\mathbf{c}, \mathbf{x}) : \mathcal{S}, \mathcal{O} \to [0, 1]$, are found by simply multiplying the upper or lower bounds of the likelihood and prior probabilities together:

$$p_U(\mathbf{c}, \mathbf{x}) = l_U(\mathbf{c}, \mathbf{x}) r_U(\mathbf{c}),$$
$$p_L(\mathbf{c}, \mathbf{x}) = l_L(\mathbf{c}, \mathbf{x}) r_L(\mathbf{c}).$$

This bound assumes there is a q_l probability that the likelihood lies in its q_l-percentile interval $[l_L(\mathbf{c}, \mathbf{x}), l_U(\mathbf{c}, \mathbf{x})]$ and, likewise, that there is a q_r probability that the prior lies in its q_r-percentile interval $[r_L(\mathbf{c}, \mathbf{x}), r_U(\mathbf{c}, \mathbf{x})]$. Thus, the interval $[p_U(\mathbf{c}, \mathbf{x}), p_L(\mathbf{c}, \mathbf{x})]$ defines the q-percentile confidence interval for posterior $\Pr(\mathbf{c}|\mathbf{x})$ where $q = q_l q_r$. In other words, there is a q probability that the posterior $\Pr(\mathbf{c}|\mathbf{x})$ lies in bounds $[p_U(\mathbf{c}, \mathbf{x}), p_L(\mathbf{c}, \mathbf{x})]$.

3.4 Narrowing the Space of Possible Causes

We have established $\mathcal{S}$ as the *infinite* space containing all possible likelihood distributions over $\mathcal{O}$ and, thus, the space of *all* possible causes. However, this space includes likelihood distributions generated by causes that are implausible, purely hypothetical, and/or require multiple preceding events. In the real world, we often choose the most likely cause from a smaller set of plausible causes; for example, one would not consider an atomic bomb to be a plausible cause for your headache disappearing. Rather than considering the entirety of $\mathcal{S}$ as the pool of possible causes, we assume that some finite subset $\mathcal{C} \subset \mathcal{S}$ with cardinality $k = |\mathcal{C}|$ has been pre-selected as the finite set of *plausible* causes assumed to contain the true cause. We further assume $\mathcal{C}$ includes a "cause" C_{other}, whose posterior encapsulates the (likely low) combined probability of all other causes in $\mathcal{S}$ occurring. With this, we assume that all causes in $\mathcal{C}$ are disjoint[2] and that $\mathcal{C}$ contains the *one* true explanation for observation $\mathbf{x}$.

[2]In other words, one event/cause in $\mathcal{C}$ cannot cause another event in $\mathcal{C}$ and any two events cannot occur simultaneously.

Definition 3.2. ($\mathcal{C}$) Let $\mathcal{C} \subset \mathcal{S}$ denote the relevant finite subset of possible cause vectors in $\mathcal{S}$.

For notational simplicity, we additionally denote each cause as $C_i \in \mathcal{C}$ and its corresponding "true" posterior probability as M_i in posterior set $\mathcal{M}$. We likewise simplify the notation of the upper and lower bounds of q-percentile uncertainty interval posterior M_i as follows: from $p_U(\mathbf{c}, \mathbf{x})$ and $p_L(\mathbf{c}, \mathbf{x})$ to u_i and l_i, respectively. For future reference, we define the following:

Definition 3.3. (M_i) Let $M_i \in \mathcal{M}$ denote the "true" posterior probability of cause $C_i \in \mathcal{C}$, where $M_i = \Pr(C_i|\mathbf{x}) \Pr(C_i)$. Then M_i falls into the following uncertainty interval with probability q:

$$M_i \in [l_i, u_i].$$

Since we assume each $C_i \in \mathcal{C}$ is disjoint, and that $\mathcal{C}$ surely contains the true explanation for observation $\mathbf{x}$, each posterior probability $\Pr(C_i|\mathbf{x})$ sums to 1. Thus,

$$\sum_{M_i \in \mathcal{M}} M_i = \Pr(\mathbf{x}).$$

Definition 3.4. ($\mathcal{U}$) Let $\mathcal{U}$ denote the set containing the q-percentile uncertainty interval bounds $[l_i, u_i]$ for each posterior $M_i \in \mathcal{M}$.

Note that we also assume $|\mathcal{M}| = |\mathcal{C}| = |\mathcal{U}| \geq 2$, as determining the most likely cause from a set of only one is trivial.

4 Abduction by Bayesian Inference

4.1 Cause Selection with Uncertainty Intervals

Given the set of q-percentile confidence posterior probability uncertainty bounds $[l_i, u_i] \in \mathcal{U}$ for each cause $C_i \in \mathcal{C}$, one selects the cause whose *point estimate posterior probability* is highest. Since the true posterior probabilities of each cause are unknown, this process may incorrectly select a cause whose posterior is not the true maximum. We quantify this rate of incorrect selection in the case where every posterior $M_i \in \mathcal{M}$ is contained in respective confidence bound $[l_i, u_i]$. Let predicate $\text{IsMax}(M_i)$ denote whether posterior M_i is truly the highest posterior. We first define the probability range where the maximum posterior must lie, $[l, u]$.

Definition 4.1. Let each posterior $M_i \in \mathcal{M}$ occur within q-percentile confidence interval $[l_i, u_i] \in \mathcal{U}$. Then, we set

$$l = \max(\{l_i | i \in \mathbb{Z}_+, i \leq |M|\})$$

and

$$u = \max(\{u_i | i \in \mathbb{Z}_+, i \leq |M|\}).$$

Proposition 4.1. *Assuming that every $M_i \in \mathcal{M}$ lies in respective q-percentile confidence interval $[l_i, u_i] \in \mathcal{U}$, the max posterior is bounded by u and l.*

Thus, in the case that *every* confidence bound fully contains its respective posterior almost surely (instead of just with probability q), any posterior M_i whose uncertainty bounds $[l_i, u_i]$ overlap with $[l, u]$ is potentially the maximum posterior with some probability $\Pr(\text{IsMax}(M_i))$.

Theorem 4.2. *Let $\mathcal{M}' \subseteq \mathcal{M}$ denote the set of posteriors whose confidence intervals intersect with $[l, u]$. The probability that $M_i \in \mathcal{M}'$ is the maximum posterior is as follows:*

$$\Pr(\mathit{IsMax}(M_i)) = \int_l^u \Pr\left(M_i = x, \bigcap_{\substack{M_j \in \mathcal{M}', \\ C_j \neq C_i}} (M_j < x) \right) dx.$$

This accounts for any estimated posterior probability distribution within $[l_i, u_i]$, but assumes M_i is contained by $[l_i, u_i]$ with probability 1.

4.2 Bayes Error Rate

However, even assuming the cause with the true highest posterior is successfully identified, there is the unavoidable error from non-zero posteriors of the "losing" categories. The true cause of a feature may simply not have the highest posterior. This minimum achievable error is expressed by Bayes Error Rate (BER):

Definition 4.3. ϵ, (Sekeh et al., 2020))

Let ϵ denote Bayes multiclass error rate (BER) for every $C_i \in \mathcal{C}$. For $|\mathcal{C}| = k$ possible causes:

$$\epsilon = 1 - \int \max\{\Pr(C_1)\Pr(\mathbf{x}|C_1), ..., \Pr(C_k)\Pr(\mathbf{x}|C_k)\}d\mathbf{x},$$

which simplifies to

$$\epsilon = 1 - \int \Pr(\mathbf{x}) \max_i \Pr(C_i|\mathbf{x})d\mathbf{x}.$$

However, the formula above is often impractical to compute for $k > 2$ causes. Instead, one can derive bounds for the multi-cause BER with techniques such as the Bhattacharyya bound, estimations using Friedman-Rafsky test statistics, and non-parametric bounds using Henze-Penrose divergence (Sekeh et al., 2018). We adopt a recent method[3] of upper bounding BER through global minimal spanning trees (Sekeh et al., 2020) and adopt a pairwise computational lower bounding method for BER (Lin, 1991).

Definition 4.4. (ϵ_{upper}, (Sekeh et al., 2020)) Let ϵ_{upper} denote the upper bound of BER such that $\epsilon \leq \epsilon_{\text{upper}}$. Then, for $|\mathcal{C}| = k$,

$$\epsilon_{\text{upper}} = 2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} \delta_{ij}$$

where

$$\delta_{ij} := \int \frac{\Pr(C_i)\Pr(C_j)\Pr(\mathbf{x}|C_i)\Pr(\mathbf{x}|C_j)}{\Pr(C_i)\Pr(\mathbf{x}|C_i) + \Pr(C_j)\Pr(\mathbf{x}|C_j)}d\mathbf{x}.$$

Definition 4.5. (ϵ_{lower}, (Wisler et al., 2016), (Lin, 1991)) Let ϵ_{lower} denote the lower bound of BER such that $\epsilon \geq \epsilon_{\text{lower}}$. BER may be lower bounded by applying pairwise computations of Bayes error ϵ_{ij} for i and j between every unique cause pair (C_i, C_j) where $C_i \in \mathcal{C}, C_i \in \mathcal{C}, i \neq j$:

$$\epsilon_{lower} = \frac{2}{k}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} (\Pr(C_i) + \Pr(C_j))\epsilon_{ij}.$$

4.3 Abductive Error Guarantees

Assume an algorithm selects from the set of possible causes $\mathcal{C}$ the cause with the highest estimated posterior. The preceding subsections detail the two possible sources of error:

1. Incomplete or imprecise background information (e.g., not knowing all the potential causes and causal relationships). This uncertainty is represented through q-percentile posterior confidence intervals in $\mathcal{U}$.

2. The true cause is not the cause with the highest true posterior. If the exact likelihood and prior is given, this minimum achievable error is simply expressed through the Bayes Error Rate (Definition 4.2.1).

We derive bounds of the error rate by combining these two possible sources of error. Let W denote the event of incorrect abduction (not selecting the true cause). Then, the probability of correctly selecting the

[3] This method provides a tighter bound than aforementioned techniques (Sekeh et al., 2020).

maximum posterior M_i and incorrect abduction is

$$\begin{aligned}\Pr(\mathrm{W}, \mathrm{IsMax}(M_i)) &= \Pr(\mathrm{W}|\mathrm{IsMax}(M_i))\Pr(\mathrm{IsMax}(M_i)) \\ &= \epsilon \Pr(\mathrm{IsMax}(M_i)).\end{aligned}$$

The probability of both incorrectly selecting the maximum posterior and incorrect abduction is

$$\begin{aligned}\Pr(\mathrm{W}, \neg\mathrm{IsMax}(M_i)) &= \Pr(\mathrm{W}|\neg\mathrm{IsMax}(M_i))\Pr(\neg\mathrm{IsMax}(M_i)) \\ &= (1 - \Pr(M_i|\mathbf{x}))(1 - \Pr(\mathrm{IsMax}(M_i))).\end{aligned}$$

Such definitions let us derive upper and lower bounds for the error rate assuming that all posteriors $M_i \in \mathcal{M}$ lie in q-percentile confidence intervals $[l_i, u_i] \in \mathcal{U}$ with probability 1. Let γ_i denote the error rate given this assumption.

Theorem 4.6. *Let γ_i denote the error rate of selected cause C_i when assuming posterior M_i lies in confidence interval $[l_i, u_i]$ almost surely. Then, γ_i is bounded above by*

$$\gamma_i \leq \epsilon_{upper} \Pr(\mathit{IsMax}(M_i)) + (1 - l_i)(1 - \Pr(\mathit{IsMax}(M_i)))$$

where ϵ_{upper} may be derived by Definition 4.2.2.

Theorem 4.7. *Let γ_i denote the error rate of selected cause C_i when assuming posterior M_i lies in confidence interval $[l_i, u_i]$ almost surely. Then, γ_i is bounded below by*

$$\gamma_i \geq \epsilon_{lower} \Pr(\mathit{IsMax}(M_i)) + (1 - u_i)(1 - \Pr(\mathit{IsMax}(M_i)))$$

where ϵ_{lower} may be derived by Definition 4.2.3.

We extend this result to the general case where all posteriors $M_i \in \mathcal{M}$ are assumed to jointly lie in their respective confidence intervals $[l_i, u_i] \in \mathcal{U}$ with probability q.

Theorem 4.8. *Let q^k be the probability that all $M_i \in \mathcal{M}$ lie in their respective confidence bounds $[l_i, u_i] \in \mathcal{U}$. Let $\gamma_{i,\ upper}$ be the upper bound of γ_i defined in Theorem 4.6. Then, the upper bound of the general error rate is given by*

$$\Pr(W) \leq 1 - q^k(1 - \gamma_{i,\ upper}).$$

Theorem 4.9. *Let q^k be the probability that all $M_i \in \mathcal{M}$ lie in their respective confidence bounds $[l_i, u_i] \in \mathcal{U}$. Let $\gamma_{i,\ lower}$ be the lower bound of γ_i defined in Theorem 4.7. Then, the lower bound of the general error rate is given by*

$$\Pr(W) \geq \gamma_{i,\ lower} q^k.$$

We note that the upper bound $1 - q^k(1 - \gamma_{i,upper}) < 1$ and the lower bound $\gamma_{i,lower} q^k > 0$, so our bounds for $\Pr(W)$ are nontrivial, being strictly tighter than the general bounds on probabilities (e.g., $[0, 1]$).

We should note that the bounds presented in this section assume noiseless observations. That is, we assume observation $\mathbf{x}$ is a wholly accurate description of the "true" outcomes of a cause. A noisy observation vector may have entries that deviate from the "true" outcome of a cause, akin to the possibility of a faulty observer or inaccurate data pipeline with which observations is processed (i.e., faulty equipment, random errors in sampling, etc.). Accounting for noisy observations for selecting the highest posterior cause is a subject of future work, and may involve the averaging of posteriors among a probability distribution of observation vectors.

The next section explores a different set of bounds describing the selection of *any* cause whose probability is above some threshold. With this broader definition of "success," we can account for noisy observations through applying the Algorithmic Search Framework (Montañez, 2017).

5 Search and Heuristic Applications

The Algorithmic Search Framework (ASF) characterizes learning problems as search, allowing one to equate the chance of success of any learning algorithm to that of a search process described by the three-tuple (Ω, T, F) – the *search space*, *target set*, and *external information resource*, respectively (Montañez, 2017). We have previously discussed abductive success in terms of finding the one "true" cause for some observation vector (which may or may not have the highest posterior) *assuming* the selection of the single highest posterior. Furthermore, we assumed noiseless observations. By reframing the ASF for abduction, we describe an algorithm's ability to identify the cause(s) with posteriors above some threshold in terms of information-theoretic properties within (Ω, T, F) and generalize to noisy observation vectors.

5.1 ASF: Success of Abduction through Search

We define each term of (Ω, T, F) as follows.

Search Space (Ω) constitutes the finite set of pre-selected, plausible causes for the given observation vector $\mathbf{x}$; it is synonymous with $\mathcal{C}$ in Definition 3.4.1. P_i over search space Ω denotes the probability distribution over the space at step i, and $P_i(T)$ is the probability of success – namely, the amount of probability mass placed on the target set T at time i (Montañez, 2017). In our adaptation, P_i denotes the posterior distribution of $\Pr(C_i|\mathbf{x})$ over all possible causes C_i in Ω. P_i may be derived from aforementioned bounds $[l_i, u_i] \in \mathcal{U}$ of posterior-adjacent value $\Pr(C_i|\mathbf{x})\Pr(\mathbf{x})$ (Definition 4.1.1) with two modifications: (1) P_i denotes the *point estimate probability* of the posterior within these confidence bounds, and (2) this point estimate of $\Pr(C_i|\mathbf{x})\Pr(\mathbf{x})$ is inversely scaled by $\Pr(\mathbf{x})$ such that P_i is a valid probability mass function that sums to one.

Target Set (T), a subset of the search space Ω, contains the set of the "more plausible" causes with posterior probability P_i above or at *minimum performance value* in $(0, 1]$. Search aims to identify causes in Ω that lie in T, a task whose difficulty increases as the threshold for T rises.

External Information Resource (F) embeds (1) the observation vector $\mathbf{x}$ whose cause we determine, and (2) the upper and lower bounds of the q-confidence intervals for likelihood and prior probabilities across Ω for every cause $C_i \in \Omega$. More specifically, F contains the likelihood bounds $l_U(\mathbf{c}, \mathbf{x})$ and $l_L(\mathbf{c}, \mathbf{x})$ and prior bounds $r_U(\mathbf{c}, \mathbf{x})$ and $r_L(\mathbf{c}, \mathbf{x})$, which inform the construction posterior probability distribution P_i over Ω for the search process as defined previously. Note that, as explained in Section 4, the ASF places few restrictions on information resources F, and thus allows for both noisy or noiseless observations.

Framing abduction through the ASF, we apply established derivations of the maximal success probability of success defined in terms of information-theoretic properties of (Ω, T, F) and the complexity of the search problem (Montañez, 2017).

Theorem 5.1. *The probability of a successful abduction, q, is bounded above by*

$$q \leq \frac{I(T;F) + D(P_T||\mathcal{U}_T) + 1}{I_\Omega},$$

where $I_\Omega = -\log \frac{|T|}{|\Omega|}$, $D(P_T||\mathcal{U}_T)$ is the Kullback-Leibler divergence between the marginal distribution on target sets and the uniform distribution on possible target sets, and $I(T;F)$ is the mutual information between the target and observation.

We interpret $I(T;F)$ as the dependence between the target set and the observation, $D(P_T||\mathcal{U}_T)$ as the non-uniformness of the target, and I_Ω as the sparseness of the targets inside the search space. When the true cause is highly correlated with the observations (i.e., less random), the achievable success rate is high. When the search space consists of a large number of causes, the achievable success rate is lower. This gives us an additional information-theoretic upper bound on the probability of successful abduction.

5.2 ASF: High-Likelihood Causes are Rare

Any high-posterior cause must also confer high-likelihood to observed effects, due to the multiplicative nature of posterior computation. Yet a cause can only make an observation vector more probable at the cost of making others less probable. Such high-likelihood causes must necessarily be rare to the degree they confer high joint-probability on the observations, as shown by the following theorem (Montañez, 2017).

Theorem 5.2. *(Famine of Favorable Strategies Theorem, (Montañez, 2017)) For any fixed search problem* (Ω, T, F)*, set of probability mass functions* $\mathcal{P} = \{P : P \in [0,1]^{|\Omega|}, \sum_j P_j = 1\}$*, and a fixed threshold* $q_{\min} \in [0,1]$*,*

$$\frac{\mu(\mathcal{G}_{t,q_{\min}})}{\mu(\mathcal{G}_{\mathcal{P}})} \leq \frac{p}{q_{\min}},$$

where $p = \frac{|T|}{|\Omega|}$*,*$\mathcal{G}_{t,q_{\min}} = \{P : P \in \mathcal{P}, t^\top P \geq q_{\min}\}$*, and* μ *is Lebesgue measure.*

In contrast to Section 5.1, we consider a different search problem in applying Theorem 5.2. The search space Ω no longer consists of posteriors, but is now the space of all possible observation vectors, some of which are "close enough" to the true vector to comprise a noisy target set, T. Causes sample observation vectors by producing effects: a blind, weighted search. F becomes irrelevant. Theorem 5.2 then tells us that the proportion of causes which confer at least $q_{\min}$ probability to the observation set is necessarily small whenever $q_{\min}$ is high, if we are only willing to tolerate so much noise in our observations (leading to small $|T|$).

One might argue that although not many causes can confer high *joint* likelihood to the observations, several independent causes might together constitute an abductive explanation for the observed phenomena, if each sufficiently raises the likelihood of a *single* observed feature. Simple arithmetic renders this possibility unpersuasive. Assuming independent causes for each observed feature, the probability of jointly occurring outcomes in an observation vector $\mathbf{x}$ scales exponentially with $|\mathbf{x}|$ or the number of features. For instance, if two features have a 50/50 chance of occurring coincidentally, then the chance of them occurring together is $1/2 \cdot 1/2 = 1/4$. For four such features, the probability drops to 6.25%. Thus, the coincidental co-occurrence of independent causes that together explain an observation vector is unlikely as the number of observations increases.

5.3 Increasing Certainty in Abductive Inference

Inductive inference error guarantees derive their strength from data abundance: increasing the number of observed examples typically increases the tightness of such bounds. In contrast, abductive inference proceeds from a single observation. How do we increase confidence in our abductive judgment? In the real world, our confidence in abductive reasoning typically depends on the amount of evidence supporting or contradicting a potential hypothesis. Though consisting of a single example, there are often many features of that observation, which may or may not be well-explained by a proposed cause. This suggests a "horizontal" mode of confirmation built on many conditionally independent features, rather than the "vertical" mode of confirmation based on many observed examples typical of inductive inference. We note the importance of conditional independence among features, since features that necessarily imply each other even given the cause do not give us additional confidence in our abductive judgment.

Recall that observation vector $\mathbf{x} \in \mathcal{O}$ consists of binary features representing the existence or non-existence of some conditionally independent observed outcome. Letting $x_1, \ldots, x_n$ represent each feature of $\mathbf{x} \in \mathcal{O}$ where $|\mathbf{x}| = \dim(O) = n$, we quantitatively demonstrate this phenomenon with the following theorem.

Theorem 5.3. *For each conditionally independent feature* $x_1, \ldots, x_n$*, define* $\beta_i > 0$ *such that for all* $i = 1...n$*,*

$$\Pr(x_i|C) = \beta_i \Pr(x_i|\overline{C}).$$

Let $\beta = \sqrt[n]{\prod_i^n \beta_i}$*, the geometric mean of the* β_i*. If* $\beta > 1$*, then*

$$\lim_{n\to\infty} \frac{\Pr(x_1, \ldots, x_n|C)}{\Pr(x_1, \ldots, x_n|\overline{C})} = \lim_{n\to\infty} \beta^n = \infty.$$

Each conditionally independent observation feature can either support ($\beta_i > 1$) or contradict ($\beta_i < 1$) the proposed cause. If features support the current cause C on average (i.e., $\beta > 1$), then the confidence of abduction (ratio between likelihood under C over $\overline{C}$) approaches infinity as the number of (on average) supporting features increases.

6 Discussion

We formalize abduction as selecting the cause with the highest estimated posterior from some finite pool of causes. Our focus on single-cause abduction problems is justified by their foundational role in simplifying complex decision-making processes, allowing for more precise modeling and analysis that lays the groundwork for tackling multi-cause scenarios with greater accuracy in future research. For k possible causes whose posteriors are estimated within a confidence interval set with joint probability q^k, the probability of incorrect abduction $\Pr(W)$ is bounded below by $\Pr(W) \geq \gamma_{i,lower} q^k$ (Theorem 4.9) and bounded above by $\Pr(W) \leq 1 - q^k(1 - \gamma_{\text{i, upper}})$ (Theorem 4.8). As q approaches 1, the bounds on the error rate depend more heavily on γ_i (Theorems 4.6, 4.7), which scales with the Bayes Error Rate and the amount of overlap between uncertainty intervals. One should obtain comprehensive and representative training data (i.e., maximizing q) to achieve better estimates of posteriors and thus minimize error.

Extending this formalization to the ASF, we re-frame abductive success in information-theoretic terms and account for noisy observations. In this case, the maximum success rate of abduction is governed by the complexity of the search problem and other information-theoretic properties (Theorem 5.1). The maximum success rate increases as the plausible causes (i.e., causes whose posteriors are above the *minimum performance value*) become more explainable and less random; inherent unpredictability is brought from the randomness of the "true" cause. However, one may constrain this randomness by decreasing the sparseness of the search space and/or excluding less probable causes.

Regarding the practicality of our results, it has been shown that the bounds on the Bayes Error Rate can be empirically estimated by learning from training data instead of density estimation (Sekeh et al., 2020). Then, the complexity of our bounds on abductive inference depends on the estimation of the posterior distributions. In practice, it is possible to model a selective abduction problem using a Bayesian Neural Network and obtain posterior distributions. However, it is still computationally expensive to compute the integrals and would only be practical to estimate in small neural networks numerically.

The mathematical formalization and bounds established in our paper have implications for human-like reasoning abilities which are crucial for understanding the limits of decision-making processes in artificial intelligence. Theorem 5.2 demonstrates how high-likelihood causes are rare; one is less likely to stumble across them accidentally. In addition, more supporting observations increase our confidence in a unified causal explanation, instead of the coincidental co-occurrence of observed effects. Furthermore, Theorem 5.3 aims to capture the degree of certainty of our everyday abductive inferences. Consider a scenario where we are trying to convict a suspect of a crime. If pieces of evidence collectively support that the victim is guilty, our confidence to convict grows as the amount of such evidence grows. However, if pieces of evidence were heavily contradictory and/or refuted a suspect's involvement, then we become less confident of a conviction. Our confidence would approach 0 as the amount of (on average) contradictory observations tends toward infinity.

7 Conclusion

Abductive reasoning is a key component of human rationality and discovery. State-of-the-art artificial intelligence is currently incapable of performing abductive reasoning at a human level. To achieve true human-like reasoning, it is important to consider the process of abduction and incorporate such ability in future developments.

Our work formalizes selective abduction, deriving formal error guarantees for abductive reasoning within a finite space of causes. Also, by viewing selective abduction through the lens of the Algorithmic Search Framework, we better understood how the inherent complexities of abductive inference problems affect the achievable success rates. Future work might explore creative abduction using our framework as a starting point. Creative abduction can be represented through a search space that is potentially infinite and continuous. Statistical bounds within such a framework would hold implications for general scientific reasoning, discovery, and human creativity.

References

Marco Alberti, Federico Chesani, Marco Gavanelli, Evelina Lamma, Paola Mello, and Paolo Torroni. Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference. *ACM Transactions on Computational Logic (TOCL)*, 9(4):1–43, 2008.

Dean Allemang, Michael C Tanner, Tom Bylander, and John R Josephson. Computational Complexity of Hypothesis Assembly. In *IJCAI, International Joint Conference on Artificial Intelligence*, volume 87, pp. 1112–1117, 1987.

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48:1–2, 2020.

Roobaea Alroobaea, Saeed Rubaiee, Sami Bourouis, Nizar Bouguila, and Abdulmajeed Alsufyani. Bayesian Inference Framework for Bounded Generalized Gaussian-Based Mixture Model and Its Application to Biomedical Images Classification. *International Journal of Imaging Systems and Technology*, 30(1):18–30, 2020.

Ignacio Alvarez, Jarad Niemi, and Matt Simpson. Bayesian Inference for a Covariance Matrix. *arXiv preprint arXiv:1408.4050*, pp. 1–3, 2014.

Dawn An, Joo-Ho Choi, and Nam H. Kim. Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference. *Structural Health Monitoring*, 11(3):293–303, 2012. doi: 10.1177/1475921711424520. URL `https://doi.org/10.1177/1475921711424520`.

Jean-Yves Audibert. Fast Learning Rates in Statistical Inference through Aggregation. *The Annals of Statistics*, 37(4):1591 – 1646, 2009. doi: 10.1214/08-AOS623. URL `https://doi.org/10.1214/08-AOS623`.

F. Bergadano, V. Cutello, and D. Gunetti. *Abduction in Machine Learning*, pp. 197–229. Springer Netherlands, Dordrecht, 2000. ISBN 978-94-017-1733-5. doi: 10.1007/978-94-017-1733-5_5. URL `https://doi.org/10.1007/978-94-017-1733-5_5`.

James O Berger. *Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference*, pp. 1–45, 74–117. Springer Science & Business Media, 2013.

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*, pp. 1–32. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

M. D. Buhmann, Prem Melville, Vikas Sindhwani, Novi Quadrianto, Wray L. Buntine, Luís Torgo, Xinhua Zhang, Peter Stone, Jan Struyf, Hendrik Blockeel, Kurt Driessens, Risto Miikkulainen, Eric Wiewiora, Jan Peters, Russ Tedrake, Nicholas Roy, Jun Morimoto, Peter A. Flach, and Johannes Fürnkranz. Relational Learning. In *Encyclopedia of Machine Learning*, pp. 851–857. Springer US, 2011. doi: 10.1007/978-0-387-30164-8_719. URL `https://doi.org/10.1007/978-0-387-30164-8_719`.

Badr-Eddine Chérief-Abdellatif, Pierre Alquier, and Mohammad Emtiyaz Khan. A Generalization Bound for Online Variational Inference. In *Asian conference on machine learning*, pp. 662–677. PMLR, 2019.

Siddhartha Chib. Calculating Posterior Distributions and Modal Estimates in Markov Mixture Models. *Journal of Econometrics*, 75(1):79–97, 1996. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(95)01770-4. URL `https://www.sciencedirect.com/science/article/pii/0304407695017704`.

Romain Cosentino, Randall Balestriero, Richard Baranuik, and Behnaam Aazhang. Deep Autoencoders: From Understanding to Generalization Guarantees. In *Mathematical and Scientific Machine Learning*, pp. 197–222. PMLR, 2022.

Dennis D Cox. An Analysis of Bayesian Inference for Nonparametric Regression. *The Annals of Statistics*, pp. 903–923, 1993.

Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive Learning with Robust Generalization Guarantees. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 772–814. PMLR, 23–26 Jun 2016. URL `https://proceedings.mlr.press/v49/cummings16.html`.

A.P. Dempster. Upper and Lower Probabilities Induced by a Multivalued Mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967. ISSN 00034851. URL `http://www.jstor.org/stable/2239146`.

A.P. Dempster. A Generalization of Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):205–247, 1968. ISSN 00359246. URL `http://www.jstor.org/stable/2984504`.

Thomas G Dietterich. Limitations on Inductive Learningg. In *Proceedings of the Sixth International Workshop on Machine Learning*, pp. 124–128. Elsevier, 1989.

Thomas S Ferguson, Eswar G Phadia, and Ram C Tiwari. Bayesian Nonparametric Inference. *Lecture Notes-Monograph Series*, 17:127–150, 1992.

Saurabh Garg, Sivaraman Balakrishnan, Zico Kolter, and Zachary Lipton. Ratt: Leveraging Unlabeled Data to Guarantee Generalization. In *International Conference on Machine Learning*, pp. 3598–3609. PMLR, 2021.

A. C. Kakas, R. A. Kowalski, and F. Toni. Abductive Logic Programming. *Journal of Logic and Computation*, 2(6):719–770, 12 1992. ISSN 0955-792X. doi: 10.1093/logcom/2.6.719. URL `https://doi.org/10.1093/logcom/2.6.719`.

Jörg-Uwe Kietz. Some Lower Bounds for the Computational Complexity of Inductive Logic Programming. In *Machine Learning: ECML-93*, Lecture notes in computer science, pp. 115–123. Springer Berlin Heidelberg, Berlin, Heidelberg, 1993.

Erik J Larson. *The Myth of Artificial Intelligence*, pp. 89–234. Harvard University Press, London, England, April 2021.

Hector J. Levesque. A Knowledge-Level Account of Abduction. In *IJCAI, International Joint Conference on Artificial Intelligence*, volume 11, pp. 1061–1067, 1989.

Richard A. Levine and George Casella. Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–39, 2001. URL `http://www.jstor.org/stable/1391097`.

J. Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37 (1):145–151, 1991. doi: 10.1109/18.61115.

Bojan Mihaljević, Concha Bielza, and Pedro Larrañaga. Bayesian Networks for Interpretable Machine Learning and Optimization. *Neurocomputing*, 456:648–665, 2021.

George D Montañez. The Famine of Forte: Few Search Problems Greatly Favor Your Algorithm. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, pp. 477–482. IEEE, 2017.

Raymond J Mooney. Integrating Abduction and Induction in Machine Learning. *Abduction and Induction: essays on their relation and integration*, pp. 181–191, 2000.

Hwee Tou Ng and Raymond J Mooney. An Efficient First-Order Horn-Clause Abduction System Based on the ATMS. In *Proceedings of the Ninth National Conference on Artificial Intelligence. Anaheim, CA*, pp. 494–499, 1991.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational Bayes. In *International Conference on Artificial Intelligence and Statistics*, pp. 1579–1588. PMLR, 2018.

Gabriele Paul. *AI Approaches to Abduction*, pp. 35–98. Springer Netherlands, Dordrecht, 2000. ISBN 978-94-017-1733-5. doi: 10.1007/978-94-017-1733-5_2. URL https://doi.org/10.1007/978-94-017-1733-5_2.

Charles S Peirce, Morris R Cohen, and John Dewey. Deduction, Induction, and Hypothesis. In *Chance, love, and logic*, pp. 131–153. Routledge, 2017.

David Poole. Representing Diagnostic Knowledge for Probabilistic Horn Abduction. In *IJCAI*, pp. 1129–1137, 1991.

David Poole. Probabilistic Horn abduction and Bayesian networks. *Artificial Intelligence*, 64(1):81–129, 1993. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(93)90061-F. URL https://www.sciencedirect.com/science/article/pii/000437029390061F.

Andrea Rafanelli, Stefania Costantini, and Andrea Omicini. Position Paper: On the Role of Abductive Reasoning in Semantic Image Segmentation. *CEUR WORKSHOP PROCEEDINGS*, pp. 75–84, 2023. URL https://hdl.handle.net/11585/933658.

Sindhu Raghavan and Raymond J Mooney. Bayesian Abductive Logic Programs. In *Statistical relational artificial intelligence*, pp. 82–87. Citeseer, 2010.

Cassiano Terra Rodrigues. The Method of Scientific Discovery in Peirce's Philosophy: Deduction, Induction, and Abduction. *Logica Universalis*, 5:127–164, 2011.

Jan-Willem Romeijn. Abducted by Bayesians? *Journal of Applied Logic*, 11(4):430–439, 2013.

G. Schurz. Patterns of Abduction. *Synthese*, 164(2):201–234, August 2007. doi: 10.1007/s11229-007-9223-4. URL https://doi.org/10.1007/s11229-007-9223-4.

Salimeh Yasaei Sekeh, Brandon Oselio, and Alfred O Hero. Multi-Class Bayes Error Estimation With a Global Minimal Spanning Tree. In *2018 56th annual allerton conference on communication, control, and computing (allerton)*, pp. 676–681. IEEE, 2018.

Salimeh Yasaei Sekeh, Brandon Oselio, and Alfred O. Hero. Learning to Bound the Multi-Class Bayes Error. *IEEE Transactions on Signal Processing*, 68:3793–3807, 2020. doi: 10.1109/TSP.2020.2994807.

Timothy Shanahan. The First Moment of Scientific Inquiry: CS Peirce on the Logic of Abduction. *Transactions of the Charles S. Peirce Society*, 22(4):449–466, 1986.

Bofan Song, Sumsum Sunny, Shaobai Li, Keerthi Gurushanth, Pramila Mendonca, Nirza Mukhia, Sanjana Patrick, Shubha Gurudath, Subhashini Raghavan, Imchen Tsusennaro, Shirley T. Leivon, Trupti Kolur, Vivek Shetty, Vidya R. Bushan, Rohan Ramesh, Tyler Peterson, Vijay Pillai, Petra Wilder-Smith, Alben Sigamani, Amritha Suresh, moni Abraham Kuriakose, Praveen Birur, and Rongguang Liang. Bayesian Deep Learning for Reliable Oral Cancer Image Classification. *Biomed. Opt. Express*, 12(10):6422–6430, Oct 2021. doi: 10.1364/BOE.432365. URL https://opg.optica.org/boe/abstract.cfm?URI=boe-12-10-6422.

Luke Tierney. Markov Chains for Exploring Posterior Distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994. ISSN 00905364. URL http://www.jstor.org/stable/2242477.

Alan A Wisler, Visar Berisha, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. Empirically-Estimable Multi-Class Classification Bounds. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2594–2598, 2016. URL https://api.semanticscholar.org/CorpusID:10821499.

David H. Wolpert and David B. Kinney. A Stochastic Model of Mathematics and Science. *Foundations of Physics*, 54(2), April 2024. ISSN 1572-9516. doi: 10.1007/s10701-024-00755-9. URL http://dx.doi.org/10.1007/s10701-024-00755-9.

Daniel Yekutieli. Identification of Correlated Damage Parameters Under Noise and Bias Using Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 74(3):515–541, 2012. ISSN 13697412, 14679868. URL http://www.jstor.org/stable/41674641.

Chi Zhang, Baoxiong Jia, Song-Chun Zhu, and Yixin Zhu. Abstract Spatial-Temporal Reasoning via Probabilistic Abduction and Execution. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9731–9741, 2021. URL `https://api.semanticscholar.org/CorpusID:232380362`.

Appendix: Proofs

Proposition 4.1. *Assuming that every $M_i \in \mathcal{M}$ lies in respective q-percentile confidence interval $[l_i, u_i] \in \mathcal{U}$, the max posterior is bounded by u and l.*

Proof. We prove that the maximum posterior probability $M_{max} \in \mathcal{M}$ is in the interval $[l, u]$ (4.1) by way of contradiction. Assume $M_{max} \notin [l, u]$. Then, knowing that $|\mathcal{M}| \geq 2$, one of the following is true:

Case 1: $M_{max} > u$. Since u is defined as the maximum of all posterior upper bounds, we reach a contradiction if $M_{max} > u$ as M_{max} would not be in $\mathcal{M}$.

Case 2: $M_{max} < l$. If $M_{max} < l$, then since l is defined as the highest lower bound, there must exist an $M_i \in \mathcal{M}$ such that $M_i \neq M_{max}$ whose lower bound $l_i \geq l$. If this is the case, $M_i > M_{max}$ since $M_i \geq l_i$, $l_i \geq l$, and $l > M_{max}$. M_{max} would not be the maximum posterior, resulting in a contradiction.

Both possibilities result in contradiction, so $M_{max} \notin [l, u]$ is not true. Thus, the maximum posterior is bounded by u and l, or $M_{max} \in [l, u]$. □

Theorem 4.2. *Let $\mathcal{M}' \subseteq \mathcal{M}$ denote the set of posteriors whose confidence intervals intersect with $[l, u]$. The probability that $M_i \in \mathcal{M}'$ is the maximum posterior is as follows:*

$$\Pr(\mathit{IsMax}(M_i)) = \int_l^u \Pr\left(M_i = x, \bigcap_{\substack{M_j \in \mathcal{M}',\\ C_j \neq C_i}} (M_j < x)\right) dx.$$

Proof. Via Proposition 4.2.1, the max posterior is bounded by u and l. So, if x is the value of M_i and M_i is the maximum posterior, $l \leq x \leq u$. For M_i to be the maximum posterior value with a value of x, both $M_i = x$ and $M_j < x$ for all $j \neq i$. So, we express $\Pr(\text{IsMax}(M_i))$ as an integral of joint probabilities:

$$\Pr(\text{IsMax}(M_i)) = \int_l^u \Pr\left(M_i = x, \bigcap_{j=1, j\neq i}^{k} (M_j < x)\right) dx.$$

□

Theorem 4.6. *Let γ_i denote the error rate of selected cause C_i when assuming posterior M_i lies in confidence interval $[l_i, u_i]$ almost surely. Then, γ_i is bounded above by*

$$\gamma_i \leq \epsilon_{upper} \Pr(\mathit{IsMax}(M_i)) + (1 - l_i)(1 - \Pr(\mathit{IsMax}(M_i)))$$

where ϵ_{upper} may be derived by Definition 4.2.2.

Proof. Let $\Pr(W_i)$ be the probability we produce a wrong abduction given we selected cause C_i (this is synonymous with the "error rate" of selecting cause C_i). By the law of total probability,

$$\Pr(W_i) = \Pr(W_i, \text{IsMax}(M_i)) + \Pr(W_i, \neg\text{IsMax}(M_i)).$$

We first discuss the case where the maximum posterior is selected (i.e., $\text{IsMax}(M_i)$ holds). Here, $\Pr(W_i \mid \text{IsMax}(M_i))$ is given by Bayes error rate, which is upper bounded by ϵ_{upper} (Definition 4.4). Thus,

$$\begin{aligned}\Pr(W_i, \text{IsMax}(M_i)) &= \Pr(W_i|\text{IsMax}(M_i)) \Pr(\text{IsMax}(M_i)) \\ &\leq \epsilon_{upper} \Pr(\text{IsMax}(M_i)).\end{aligned}$$

In the case that the highest posterior is *not* selected, the probability that we result in a false inference is given by $1 - \Pr(C_i|\mathbf{x})$, where $\Pr(C_i|\mathbf{x})$ is the theoretical true posterior. Let l_i denote the lower bound of

$\Pr(C_i|\mathbf{x})$, which we assume to bound the true posterior almost surely. Thus,

$$\begin{aligned}\Pr(W_i, \neg\text{IsMax}(M_i)) &= \Pr(W_i \mid \neg\text{IsMax}(M_i)) \Pr(\neg\text{IsMax}(M_i)) \\ &= (1 - \Pr(C_i \mid \mathbf{x}))(1 - \Pr(\text{IsMax}(M_i))) \\ &\leq (1 - l_i)(1 - \Pr(\text{IsMax}(M_i))).\end{aligned}$$

We combine our bounds to obtain

$$\Pr(W_i) \leq \epsilon_{upper} \Pr(\text{IsMax}(M_i)) \; + (1 - l_i)(1 - \Pr(\text{IsMax}(M_i))).$$

We relabel γ_i as the error rate of selected cause C_i:

$$\gamma_i \leq \epsilon_{upper} \Pr(\text{IsMax}(M_i)) + (1 - l_i)(1 - \Pr(\text{IsMax}(M_i))).$$

□

Theorem 4.7. *Let γ_i denote the error rate of selected cause C_i when assuming posterior M_i lies in confidence interval $[l_i, u_i]$ almost surely. Then, γ_i is bounded below by*

$$\gamma_i \geq \epsilon_{lower} \Pr(\textit{IsMax}(M_i)) + (1 - u_i)(1 - \Pr(\textit{IsMax}(M_i)))$$

where ϵ_{lower} may be derived by Definition 4.2.3.

Proof. Let $\Pr(W_i)$ denote the probability we produce a wrong abduction given we selected cause C_i (this is synonymous with the "error rate" of selecting cause C_i). By the law of total probability,

$$\Pr(W_i) = \Pr(W_i, \text{IsMax}(M_i)) + \Pr(W_i, \neg\text{IsMax}(M_i)).$$

We first explore the case where the highest posterior is selected (i.e., $\text{IsMax}(M_i)$ holds). Here, $\Pr(W_i \mid \text{IsMax}(M_i))$ is given by the Bayes error, which is lower-bounded ϵ_{lower} from Definition 4.5.

$$\begin{aligned}\Pr(W_i, \text{IsMax}(M_i)) &= \Pr(W_i \mid \text{IsMax}(M_i)) \Pr(\text{IsMax}(M_i)) \\ &\geq \epsilon_{lower} \Pr(\text{IsMax}(M_i))\end{aligned}$$

If the cause we have selected does not have the maximum posterior (i.e., $\text{IsMax}(M_i)$ does not hold), the probability that we result in a false inference is given by $1 - \Pr(C_i|\mathbf{x})$, where $\Pr(C_i|\mathbf{x})$ is the theoretical true posterior. Let u_i denote the upper bound of $\Pr(C_i|\mathbf{x})$ so that $1 - u_i$ lower bounds $1 - \Pr(C_i|\mathbf{x})$. Thus, we can derive the lower bound

$$\begin{aligned}\Pr(W_i, \neg\text{IsMax}(M_i)) &= (1 - \Pr(C_i \mid \mathbf{x}))(1 - \Pr(\text{IsMax}(M_i))) \\ &\geq (1 - u_i)(1 - \Pr(\text{IsMax}(M_i))).\end{aligned}$$

We combine the lower bounds of both components to obtain

$$\begin{aligned}\Pr(W_i) &= \Pr(W_i, \text{IsMax}(M_i)) + \Pr(W_i, \neg\text{IsMax}(M_i)) \\ &\geq \epsilon_{lower} \Pr(\text{IsMax}(M_i)) \; (1 - u_i)(1 - \Pr(\text{IsMax}(M_i))).\end{aligned}$$

We then relabel γ_i for error rate $\Pr(W_i)$:

$$\gamma_i \geq \epsilon_{lower} \Pr(\text{IsMax}(M_i)) \; + (1 - u_i)(1 - \Pr(\text{IsMax}(M_i))).$$

□

Theorem 4.8. *Let q^k be the probability that all $M_i \in \mathcal{M}$ lie in their respective confidence bounds $[l_i, u_i] \in \mathcal{U}$. Let $\gamma_{i,\ upper}$ be the upper bound of γ_i defined in Theorem 4.6. Then, the upper bound of the general error rate is given by*

$$\Pr(W) \leq 1 - q^k(1 - \gamma_{i,\ upper}).$$

Proof. Let CI_i be shorthand for the posterior confidence interval $[l_i, u_i] \in \mathcal{U}$ containing posterior $M_i \in \mathcal{M}$. By the law of total probability,

$$\begin{aligned}\Pr(W) &= \Pr(W, \forall(M_i \in \mathcal{M}), M_i \in CI_i) \; + \Pr(W, \exists(M_i \in \mathcal{M}) M_i \notin CI_i) \\ &= \Pr(W \mid \forall(M_i \in \mathcal{M}) M_i \in CI_i) \; \cdot \Pr(\forall(M_i \in \mathcal{M}) M_i \in CI_i) + \\ &\quad \Pr(W \mid \exists(M_i \in \mathcal{M}) M_i \notin CI_i) \; \cdot \Pr(\exists(M_i \in \mathcal{M}) M_i \notin CI_i).\end{aligned}$$

Note that the true posterior values are fixed and not random, but their estimates and confidence intervals (based on sampled data) *are* random. Given the true posterior values, data is generated from which confidence intervals are constructed and point estimates taken. The true posterior values thus act as parameters in a parameter estimation task. Given the value of such a parameter, the probability that a generated dataset and subsequent confidence interval captures the true parameter value is q, which (by d-separation) is conditionally independent of anything else that happens in the world. Specifically, any other parameter's (i.e., posterior's) value does not affect the probability that data generated using *this* parameter's value produces a confidence interval that captures it. All that matters is the specific parameter under which the data is generated. In other words, the probability that a second dataset generated from a different posterior produces a confidence interval that captures this second parameter's true value is **independent** of the outcome of the first data generation event, once we condition on the parameter. This second parameter is indeed given, as we need it to generate the data. Therefore, assuming k confidence intervals are constructed from data conditioned on their true parameter values, the joint probability of all k posterior probabilities being captured simultaneously by their respective q-percent confidence bounds is q^k. Thus,

$$\Pr(W) = \Pr(W \mid \forall(M_i \in \mathcal{M}) M_i \in CI_i) q^k \; + \Pr(W \mid \exists(M_i \in \mathcal{M}) M_i \notin CI_i)(1 - q^k).$$

We apply $\gamma_{i,upper}$ from Theorem 4.6, which is the probability of incorrect abduction assuming that all posterior probabilities fall into their respective confidence intervals. Thus, $\Pr(W \mid \forall(M_i \in \mathcal{M}), M_i \in CI_i)$ is bounded above by $\gamma_{i,\text{upper}}$. Additionally, we simply upper bound $\Pr(W \mid \forall M_i \in \mathcal{M}, M_i \notin CI_i)$ by one. Thus, we conclude

$$\Pr(W) \leq \gamma_{i,\text{upper}} q^k + (1)(1 - q^k),$$

or equivalently

$$\Pr(W) \leq 1 - q^k(1 - \gamma_{i,\text{upper}}).$$

□

Theorem 4.9. *Let q^k be the probability that all $M_i \in \mathcal{M}$ lie in their respective confidence bounds $[l_i, u_i] \in \mathcal{U}$. Let $\gamma_{i,\ lower}$ be the lower bound of γ_i defined in Theorem 4.7. Then, the lower bound of the general error rate is given by*

$$\Pr(W) \geq \gamma_{i,\ lower} q^k.$$

Proof. Let CI_i be shorthand for the posterior confidence interval $[l_i, u_i] \in \mathcal{U}$ containing posterior $M_i \in \mathcal{M}$. By the law of total probability,

$$\begin{aligned}\Pr(W) &= \Pr(W, \forall(M_i \in \mathcal{M}) M_i \in CI_i) \; + \Pr(W, \exists(M_i \in \mathcal{M}) M_i \notin CI_i) \\ &= \Pr(W \mid \forall(M_i \in \mathcal{M}) M_i \in CI_i) \; \cdot \Pr(\forall(M_i \in \mathcal{M}) M_i \in CI_i) + \\ &\quad \Pr(W \mid \exists(M_i \in \mathcal{M}) M_i \notin CI_i) \; \cdot \Pr(\exists(M_i \in \mathcal{M}) M_i \notin CI_i).\end{aligned}$$

Recall that the number of considered causes is $|\mathcal{C}| = |\mathcal{M}| = k$. Assuming all k confidence intervals simultaneously capture their respective posterior values with joint probability q^k (see discussion in the proof for Theorem 4.8), we obtain

$$\Pr(W) = \Pr(W \mid \forall(M_i \in \mathcal{M}) M_i \in CI_i) q^k + \Pr(W \mid \exists(M_i \in \mathcal{M}) M_i \notin CI_i)(1 - q^k).$$

We apply $\gamma_{i,\text{lower}}$ from Theorem 4.7, which is the probability of incorrect abduction, assuming all posterior probabilities fall into their respective confidence intervals. Thus, $\Pr(W \mid \forall(M_i \in \mathcal{M})M_i \in CI_i)$ is bounded below by $\gamma_{i,\text{lower}}$. Additionally, we simply lower bound $\Pr(W \mid \exists(M_i \in \mathcal{M})M_i \notin CI_i)$ by zero. Thus, we conclude

$$\Pr(W) \geq \gamma_{i,\text{lower}} q^k + (0)(1-q^k),$$

or equivalently

$$\Pr(W) \geq \gamma_{i,\text{lower}} q^k.$$

□

Theorem 5.3. *For each conditionally independent feature $x_1, \ldots, x_n$, define $\beta_i > 0$ such that for all $i = 1...n$,*

$$\Pr(x_i|C) = \beta_i \Pr(x_i|\overline{C}).$$

Let $\beta = \sqrt[n]{\prod_i^n \beta_i}$, the geometric mean of the β_i. If $\beta > 1$, then

$$\lim_{n\to\infty} \frac{\Pr(x_1, \ldots, x_n|C)}{\Pr(x_1, \ldots, x_n|\overline{C})} = \lim_{n\to\infty} \beta^n = \infty.$$

Proof. If

$$\Pr(x_i|C) = \beta_i \Pr(x_i|\overline{C}),$$

then

$$\beta_i = \frac{\Pr(x_i|C)}{\Pr(x_i|\overline{C})}.$$

So, we can write

$$\prod_i^n \beta_i = \prod_i^n \frac{\Pr(x_i|C)}{\Pr(x_i|\overline{C})} = \frac{\prod_i^n \Pr(x_i|C)}{\prod_i^n \Pr(x_i|\overline{C})}.$$

Since the features are conditionally independent,

$$\prod_i^n \Pr(x_i|C) = \Pr(x_1, ..., x_n|C)$$

and

$$\prod_i^n \Pr(x_i|\bar{C}) = \Pr(x_1, ..., x_n|\bar{C}).$$

Thus,

$$\beta^n = \prod_i^n \beta_i = \frac{\Pr(x_1, ..., x_n|C)}{\Pr(x_1, ..., x_n|\bar{C})}.$$

Therefore, when $\beta > 1$,

$$\lim_{n\to\infty} \frac{\Pr(x_1, ..., x_n|C)}{\Pr(x_1, ..., x_n|\bar{C})} = \lim_{n\to\infty} \beta^n = \infty.$$

□